ISS-A-049

WaveNet 音声合成における学習データの無音区間の影響

齋藤　穣†　　間野　一則†
† 芝浦工業大学大学院理工学研究科

1. はじめに

WaveNet[1]は畳み込みニューラルネットワークの一種であり，WaveNet を用いた音声合成では従来の手法に比べてより高品質な音声を生成できる．本稿では，WaveNet 音声合成における前処理である無音区間の除去が WaveNet の学習にどの程度影響を与えるかを調査する．

2. WaveNet

WaveNet は過去の信号のサンプル系列から現在のサンプルの値を予測するニューラルネットワークであり，ある特徴量 $\mathbf{h}$ が与えられた時の波形 $\mathbf{x}$ の生成確率は次の式で表される．

$$p(\mathbf{x}|\mathbf{h}) = \prod_{t=1}^{T} p(x_t|x_1, \ldots, x_{t-1}, \mathbf{h})$$

WaveNet ボコーダ[2]は従来のボコーダの音響特徴パラメータを WaveNet の特徴量として入力し音声合成を行う．

3. 研究概要

3.1.　合成音声の評価指標

本研究では合成音声の評価に次の式で表されるスペクトル歪み尺度CD を用いた．

$$CD = \frac{10}{\ln 10} 2 \sum_{i=1}^{p} (c_i - c_i')^2$$

ただし，ここで c_i, c_i'は音声のケプストラム係数である．

3.2.　提案手法

提案手法の手順は以下の通りである．

1. 量子化ビット数 8bit のデータセットを作成する．
2. 1 のデータセットの音声の先頭と末尾の無音区間を除去したデータセットを作成する．
3. STFT によりメルスペクトログラムを得る．
4. 1，2 のデータセットを用いて WaveNet ボコーダの学習を行う．入力特徴量として 3 で計算したメルスペクトログラムを用いる．
5. 学習した WaveNet ボコーダを用いて音声を再合成する．
6. 元音声と合成音声のケプストラムを算出し，スペクトル歪み尺度により学習回数と品質の関係を評価する．

4. 実験

4.1.　実験結果

入力の元の音声のスペクトログラムを図 1 に，生成された音声のスペクトログラムを図 2,3 に示す．また，学習回数とスペクトル歪みの関係を図 4 に示す．

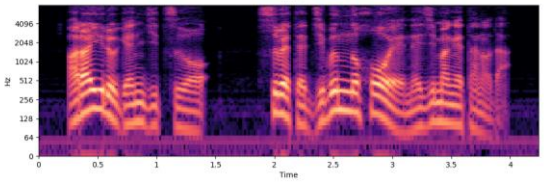

図 1　目標音声

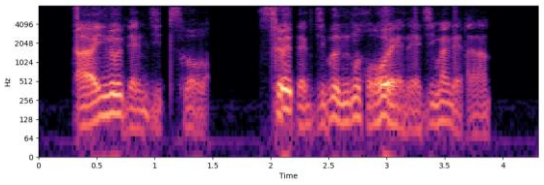

図 2　無音区間除去 生成音声

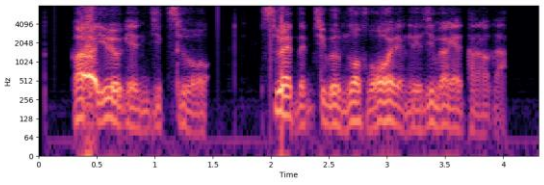

図 3　無音区間あり 生成音声

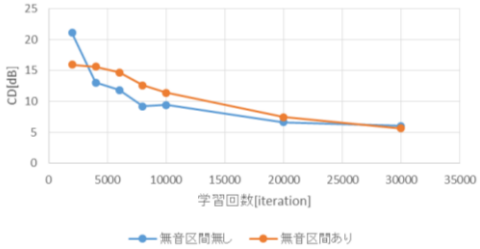

図 4　学習回数と CD の関係

4.2.　客観品質評価

客観品質評価の結果，無音区間の除去を行うと学習の進みが速くなる傾向が示されたが，学習回数を増やした時の品質の向上には寄与しないという結果が得られた．

5. まとめ

WaveNet 音声合成における無音区間が与える影響について調査し，少ない学習回数において無音区間を切り取ったデータセットを用いて学習した WaveNet ボコーダは学習が速く進むことを示した．

参考文献

[1] A. van et al., “WaveNet: A generative model for raw audio,” arXiv:1609.03499 (2016).

[2] 玉森聡ほか, “音声生成過程を考慮した WaveNet に基づく音声波形合成法,”信学技報 116(476), 1-6, (2017).